Philippe and the height of trees

N. Broutin

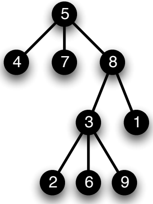
December 15, 2011

Uniform random trees from a class

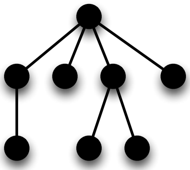
Model

- Take a class of trees
- put them all in a big bag
- pick one at random

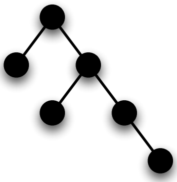
Cayley (unordered, labelled)



Plane trees (ordered children)



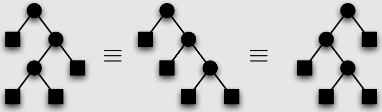
Catalan (unlabelled, positions)

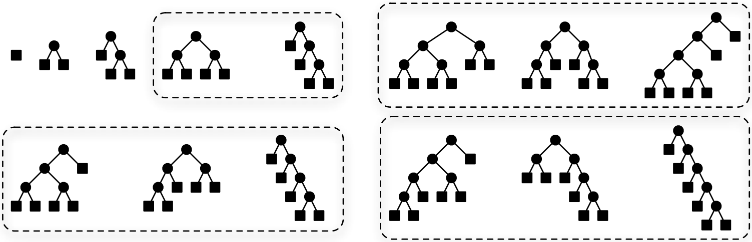


Rooted unlabelled unordered binary trees

Objects: (Pólya, Otter)

- rooted tree
- indistinguishable nodes
- (out)degree 0 or 2
- **size**: number of leaves





Uniformly random trees, why?

Applications were **originally dubious**:

- Not the trees for data structures
- Not the trees for skeleton of real-world networks
- $\Rightarrow$ not quite the model of randomness for applications

Turn out to be **essential combinatorial structures**:

- hashing with linear probing
- coalescent/fragmentation processes (additive coalescent)
- essential in the combinatorics of planar maps
- building blocks of critical random graphs
- building blocks of minimum spanning trees

and... **they are wonderful mathematical objects!**

A remarkable paper

Universality of the asymptotic behaviour of trees:

- before: limited to the cases with explicit expressions;
- demonstrate that a large number of trees behave similarly;
- implicit connection to the Brownian excursion $\Rightarrow$ CRT

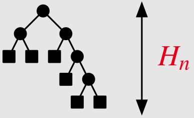
Generality of the method: **the origins of singularity analysis**

- before: estimates of series using integrals using real analysis
- estimates using the behaviour of generating function about the singularities
- gives local results, and estimates for error terms
- approach underlies the (more difficult) analysis of search trees
- has led to hundreds of beautiful results!

Model and Aim

Model of randomness:

- put all trees of size n in a bag
- take one uniformly at random
- Let H_n be the height



Proceed from first principles

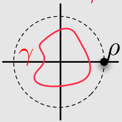
$$\mathbb{P}\left(H_n \leq h\right) = \frac{\#\{\text{trees of size } n \text{ and height } \leq h\}}{\#\{\text{trees of size } n\}}$$

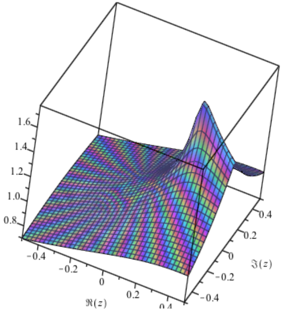
A “complex” coding-decoding problem

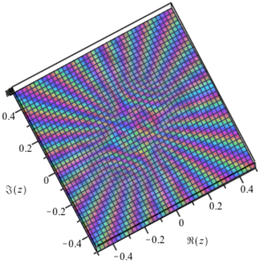
Generic decoding: Cauchy’s coefficient formula

Assume $f(z) = \sum_{n\geq 0} a_n z^n$ is analytic inside a simple contour γ around the origin,

$$a_n = \frac{1}{2i\pi} \int_\gamma f(z) \frac{dz}{z^{n+1}}.$$





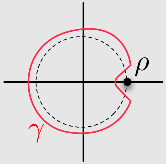


A “complex” coding-decoding problem II

Clever decoding

Suppose $f(z)$ has an analytic continuation in a “pacman” domain **beyond** the unique dominant singularity ρ

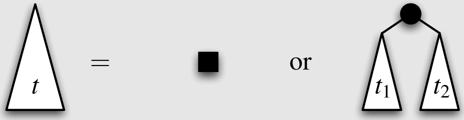
- main contribution is in the dent around ρ
- **similar functions around ρ have similar coefficients!**



A combinatorial approach

A canonical example: **binary position trees**

Recursive decomposition: $y(z) = \sum_t z^{|t|} = 1 + zy(z)^2$



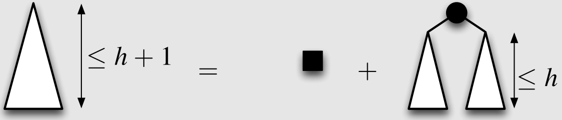
$$\sum_t z^{|t|} \quad = \quad 1 \quad + \quad z \sum_{t_1, t_2} z^{|t_1| + |t_2|}$$

Generating function: $y(z) = \sum_t z^{|t|} \equiv \sum_n y_n z^n$

$$y(z) = \frac{1 - \sqrt{1 - 4z}}{2z} \quad \text{and} \quad y_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi}} n^{-3/2} 4^n$$

An analytic-combinatorics approach

Recursive decomposition:



$$y_{h+1}(z) = 1 + zy_h(z)^2 \qquad \text{and} \qquad y_0(z) = z$$

Main recurrence: proportion of trees exceeding height h

$$e_h(z) = 1 - \frac{y_h(z)}{y(z)}$$

$$e_{h+1}(z) = zy(z)e_h(z)(2 - e_h(z)), \qquad e_0(z) = 1 - \frac{z}{y(z)}$$

Understanding H_n reduces to understanding how $y_h \to y$ or $e_h \to 0$

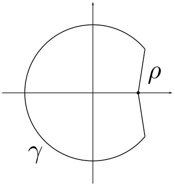
Recap what we want

H_n := height of a uniformly random tree

$$\mathbb{P}(H_n > h) = e_{h,n} \qquad \Rightarrow \qquad \textbf{estimate} \text{ the coefficients } e_{h,n}$$

Cauchy coefficient formula

$$e_{h,n} = \frac{1}{2i\pi} \int_\gamma e_h(z) \frac{dz}{z^{n+1}}$$



$\Rightarrow$ Need to **estimate** the **generating function** $e_h(z)$

- Away from ρ, large arc: **negligible (check)**
- Close to ρ, rectilinear portions **estimate precisely**

The outercircular arc: an upper bound

$$|e_{h+1}(z)| \leq |e_h(z)| \cdot \underbrace{|zy(z)|\,(2 + \overbrace{|e_h(z)|}^{\alpha})}_{\beta}$$

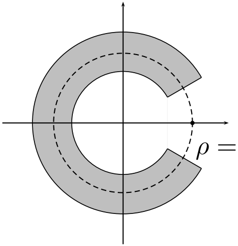
Lemma (Criterion for convergence)

Let $z \in \{z : |y| \leq 1\}$. Then

$$|e_h(z)| \to 0 \quad \Leftrightarrow \quad \exists \alpha, \beta < 1, m \in \mathbb{N} \; s.t. \begin{cases} |e_m| < \alpha \\ |zy(z)|(2+\alpha) < \beta \end{cases}$$

Proving $|e_h(z)| \to 0$ in an **extended domain**:

- $e_h \leq \sum_{n>h} y_n z^n \leq \frac{1}{\sqrt{h}} \left(\frac{z}{|\rho|}\right)^h \Rightarrow$ convergence for $|z| \leq \rho$
- continuity $\Rightarrow$ convergence inside tube: $|\arg(z)| > \eta$ and $|z| < \rho + \epsilon$



Close to the singularity: precise estimates

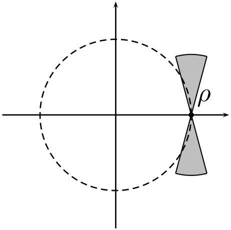
Recall $y(z) = \frac{1-\sqrt{1-4z}}{2z}$

Remark

$|e_h(z)| > 0$ and $2zy(z) \to 1$ as $z \to \rho$ $\Rightarrow$ impossible to apply the criterion directly for h **fixed**

Plan of attack: apply for $h = N(z) \to \infty$

1. control the evolution of $|e_h(z)|$ for $h \leq N(z)$ and close to the real axis: “almost” real
2. choose $N(z)$ s.t. the criterion is satisfied for $h = N(z) \Rightarrow e_h(z) \to 0$
3. bootstrap to obtain good estimates



Sketch of the approach inside the sandclock

A problem in iteration

$$e_{h+1}(z) = e_h(z) \cdot (1-\varepsilon) \cdot \left(1 - \frac{e_h(z)}{2}\right) \qquad \varepsilon(z) = \sqrt{1-4z}$$

$$\begin{aligned}\frac{(1-\varepsilon)^{h+1}}{e_{h+1}} &= \frac{(1-\varepsilon)^h}{e_h} \cdot \frac{1}{1-e_h/2} \\ &= \frac{(1-\varepsilon)^h}{e_h} + \frac{(1-\varepsilon)^h}{2} + \frac{(1-\varepsilon)^h}{4} \cdot \frac{e_h}{1-e_h/2}\end{aligned}$$

Alternative recurrence

$$\frac{(1-\varepsilon)^{h+1}}{e_{h+1}} = \frac{1}{e_0} + \sum_{i=0}^{h} \frac{(1-\varepsilon)^i}{2} + \sum_{i=0}^{h} \frac{(1-\varepsilon)^i}{4} \cdot \frac{e_i}{1-e_i/2}$$

Gathering the fruits: give me an integral!

Bootstrapping: As $z \to \rho$ we have $\varepsilon \to 0$

$$\frac{|1-\varepsilon|^h}{|e_h|} \geq \frac{K}{h} \quad \Rightarrow \quad |e_i| \leq \frac{|1-\varepsilon|^i}{Ki}$$

Main approximation in the sandclock

$$\frac{(1-\varepsilon)^h}{e_h} = \frac{1-(1-\varepsilon)^h}{2\varepsilon} + O(\log h)$$

Results

Simple family:

- c_d types of nodes of degree d, $c_d \leq \gamma^d$, for some γ
- gives a constant λ, depending on the family

Theorem (Limit distribution)

The height H_n admits a **limiting theta distribution** and $\delta^{-1}/\sqrt{\log n} \leq x \leq \delta\sqrt{\log n}$:

$$\lim_{n\to\infty}\left|\mathbb{P}\left(H_n \geq x\sqrt{n}\right) - \sum_{k\geq 1}(k^2\lambda^2x^2 - 2)e^{-k^2\lambda^2x^2/4}\right| = 0.$$

Local limit theorem

For $h = x\sqrt{n}$ an integer, $\delta^{-1}/\sqrt{\log n} \leq x \leq \delta\sqrt{\log n}$

$$\mathbb{P}\left(H_n = h\right) \sim \frac{1}{2x}\sum_{k\geq 1}(k^4\lambda^4x^4 - 6k^2\lambda^2x^2)e^{-k^2\lambda^2x^2/4}.$$

Results II

Theorem (Moments)

Let $r \geq 1$. The r-*th* moment of the height H_n satisfies

$$\mathbf{E}\left[H_n\right] \sim \frac{2}{\lambda}\sqrt{\pi n} \quad \text{and} \quad \mathbb{E}[H_n^r] \sim r(r-1)\zeta(r)\Gamma(r/2)\left(\frac{2}{\lambda}\right)^r n^{r/2}, \quad r \geq 2.$$

Some observations:

- distribution of the height of **all simply generated trees**
- distribution of the maximum of a **Brownian excursion**
- connection made explicit 10 years later by Aldous